

THE GENETIC CODE

How does the order of bases in a nucleic acid determine the order of amino acids in a protein? It seems that each amino acid is specified by a triplet of bases, and that triplets are read in simple sequence

by F. H. C. Crick

Within the past year important progress has been made in solving the "coding problem." To the biologist this is the problem of how the information carried in the genes of an organism determines the structure of proteins.

Proteins are made from 20 different kinds of small molecule—the amino acids—strung together into long polypeptide chains. Proteins often contain several hundred amino acid units linked together, and in each protein the links are arranged in a specific order that is genetically determined. A protein is therefore like a long sentence in a written language that has 20 letters.

Genes are made of quite different long-chain molecules: the nucleic acids DNA (deoxyribonucleic acid) and, in some small viruses, the closely related RNA (ribonucleic acid). It has recently been found that a special form of RNA, called messenger RNA, carries the genetic message from the gene, which is located in the nucleus of the cell, to the surrounding cytoplasm, where many of the proteins are synthesized [see "Messenger RNA," by Jerard Hurwitz and J. J. Furth; *SCIENTIFIC AMERICAN*, February].

The nucleic acids are made by joining up four kinds of nucleotide to form a polynucleotide chain. The chain provides a backbone from which four kinds of side group, known as bases, jut at regular intervals. The order of the bases, however, is not regular, and it is their precise sequence that is believed to carry the genetic message. The coding problem can thus be stated more explicitly as the problem of how the sequence of the four bases in the nucleic acid determines the sequence of the 20 amino acids in the protein.

The problem has two major aspects, one general and one specific. Specifically

one would like to know just what sequence of bases codes for each amino acid. Remarkable progress toward this goal was reported early this year by Marshall W. Nirenberg and J. Heinrich Matthaei of the National Institutes of Health and by Severo Ochoa and his colleagues at the New York University School of Medicine. [Editor's note: Brief accounts of this work appeared in "Science and the Citizen" for February and March. This article was planned as a companion to one by Nirenberg, now in preparation, which will deal with the biochemical aspects of the genetic code.]

The more general aspect of the coding problem, which will be my subject, has to do with the length of the genetic coding units, the way they are arranged in the DNA molecule and the way in which the message is read out. The experiments I shall report were performed at the Medical Research Council Laboratory of Molecular Biology in Cambridge, England. My colleagues were Mrs. Leslie Barnett, Sydney Brenner, Richard J. Watts-Tobin and, more recently, Robert Shulman.

The organism used in our work is the bacteriophage T4, a virus that infects the colon bacillus and subverts the biochemical machinery of the bacillus to make multiple copies of itself. The infective process starts when T4 injects its genetic core, consisting of a long strand of DNA, into the bacillus. In less than 20 minutes the virus DNA causes the manufacture of 100 or so copies of the complete virus particle, consisting of a DNA core and a shell containing at least six distinct protein components. In the process the bacillus is killed and the virus particles spill out. The great value of the T4 virus for genetic experiments is that many generations and billions of individuals can be produced in a short time. Colonies containing mutant indi-

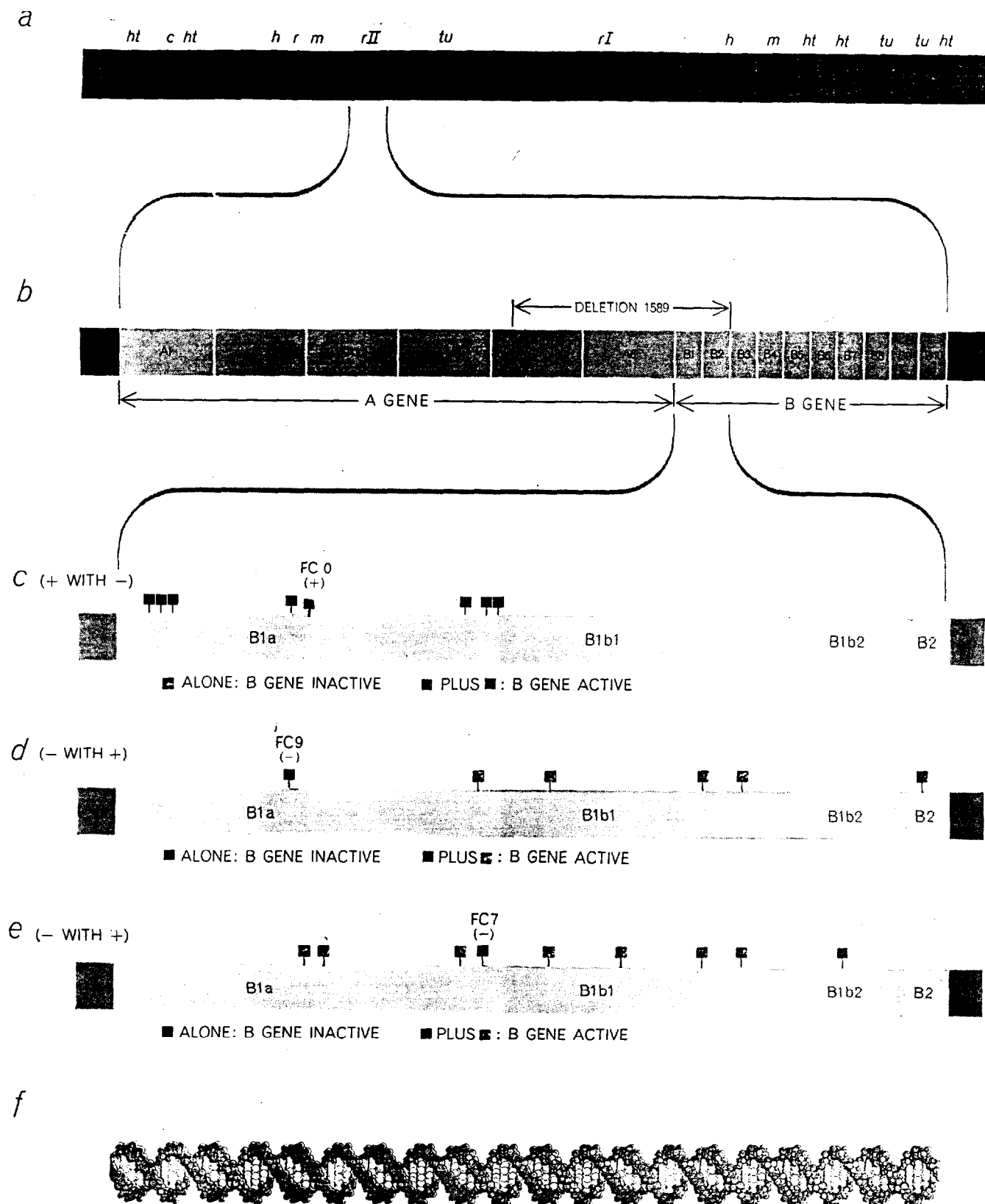
viduals can be detected by the appearance of the small circular "plaques" they form on culture plates. Moreover, by the use of suitable cultures it is possible to select a single individual of interest from a population of a billion.

Using the same general technique, Seymour Benzer of Purdue University was able to explore the fine structure of the A and B genes (or cistrons, as he prefers to call them) found at the "rII" locus of the DNA molecule of T4 [see "The Fine Structure of the Gene," by Seymour Benzer; *SCIENTIFIC AMERICAN*, January]. He showed that the A and B genes, which are next to each other on the virus chromosome, each consist of some hundreds of distinct sites arranged in linear order. This is exactly what one would expect if each gene is a segment, say 500 or 1,000 bases long, of the very long DNA molecule that forms the virus chromosome [see illustration on opposite page]. The entire DNA molecule in T4 contains about 200,000 base pairs.

The Usefulness of Mutations

From the work of Benzer and others we know that certain mutations in the A and B region made one or both genes inactive, whereas other mutations were only partially inactivating. It had also been observed that certain mutations were able to suppress the effect of harmful mutations, thereby restoring the function of one or both genes. We suspected that the various—and often puzzling—consequences of different kinds of mutation might provide a key to the nature of the genetic code.

We therefore set out to re-examine the effects of crossing T4 viruses bearing mutations at various sites. By growing two different viruses together in a common culture one can obtain "recombinants" that have some of the properties



rII REGION OF THE T4 VIRUS represents only a few per cent of the DNA (deoxyribonucleic acid) molecule that carries full instructions for creating the virus. The region consists of two genes, here called A and B. The A gene has been mapped into six major segments, the B gene into 10 (*b*). The experiments reported in this article involve mutations in the first and second segments of the B gene. The B gene is inactivated by any mutation

that adds a molecular subunit called a base (*colored square*) or removes one (*black square*). But activity is restored by simultaneous addition and removal of a base, as shown in *c*, *d* and *e*. An explanation for this recovery of activity is illustrated on page 70. The molecular representation of DNA (*f*) is estimated to be approximately in scale with the length of the B1 and B2 segments of the B gene. The two segments contain about 100 base pairs.

of one parent and some of the other. Thus one defect, such as the alteration of a base at a particular point, can be combined with a defect at another point to produce a phage with both defects [see upper illustration below]. Alternatively, if a phage has several defects, they can be separated by being crossed

with the "wild" type, which by definition has none. In short, by genetic methods one can either combine or separate different mutations, provided that they do not overlap.

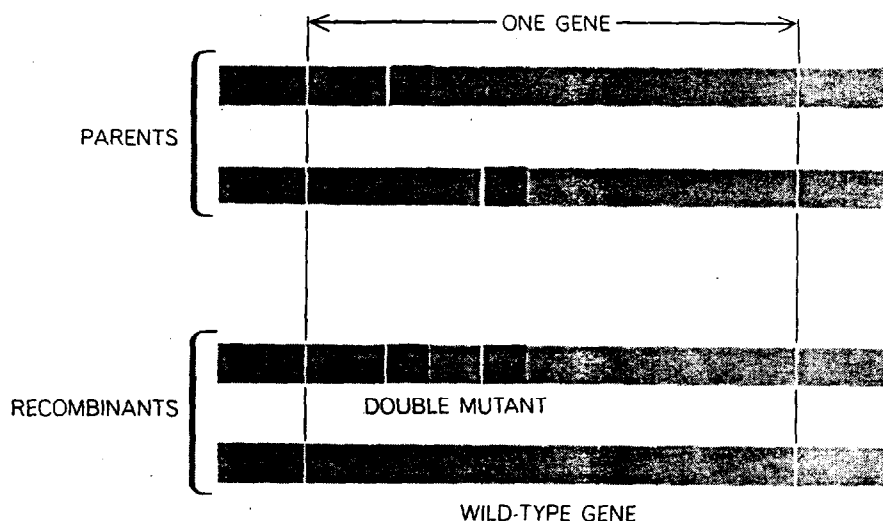
Most of the defects we shall be considering are evidently the result of adding or deleting one base or a small group

of bases in the DNA molecule and not merely the result of altering one of the bases [see lower illustration on this page]. Such additions and deletions can be produced in a random manner with the compounds called acridines, by a process that is not clearly understood. We think they are very small additions or deletions, because the altered gene seems to have lost its function completely; mutations produced by reagents capable of changing one base into another are often partly functional. Moreover, the acridine mutations cannot be reversed by such reagents (and vice versa). But our strongest reason for believing they are additions or deletions is that they can be combined in a way that suggests they have this character.

To understand this we shall have to go back to the genetic code. The simplest sort of code would be one in which a small group of bases stands for one particular acid. This group can scarcely be a pair, since this would yield only 4×4 , or 16, possibilities, and at least 20 are needed. More likely the shortest code group is a triplet, which would provide $4 \times 4 \times 4$, or 64, possibilities. A small group of bases that codes one amino acid has recently been named a codon.

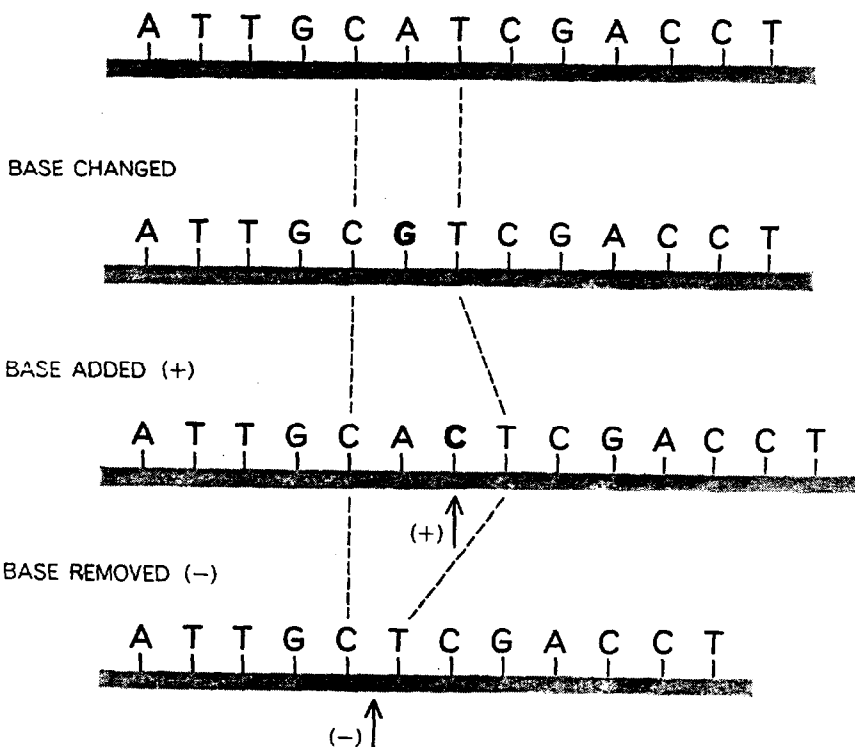
The first definite coding scheme to be proposed was put forward eight years ago by the physicist George Gamow, now at the University of Colorado. In this code adjacent codons overlap as illustrated on the opposite page. One consequence of such a code is that only certain amino acids can follow others. Another consequence is that a change in a single base leads to a change in three adjacent amino acids. Evidence gathered since Gamow advanced his ideas makes an overlapping code appear unlikely. In the first place there seems to be no restriction of amino acid sequence in any of the proteins so far examined. It has also been shown that typical mutations change only a single amino acid in the polypeptide chain of a protein. Although it is theoretically possible that the genetic code may be partly overlapping, it is more likely that adjacent codons do not overlap at all.

Since the backbone of the DNA molecule is completely regular, there is nothing to mark the code off into groups of three bases, or into groups of any other size. To solve this difficulty various ingenious solutions have been proposed. It was thought, for example, that the code might be designed in such a way that if the wrong set of triplets were chosen, the message would always be complete nonsense and no protein would



GENETIC RECOMBINATION provides the means for studying mutations. Colored squares represent mutations in the chromosome (DNA molecule) of the T4 virus. Through genetic recombination, the progeny can inherit the defects of both parents or of neither.

WILD-TYPE GENE



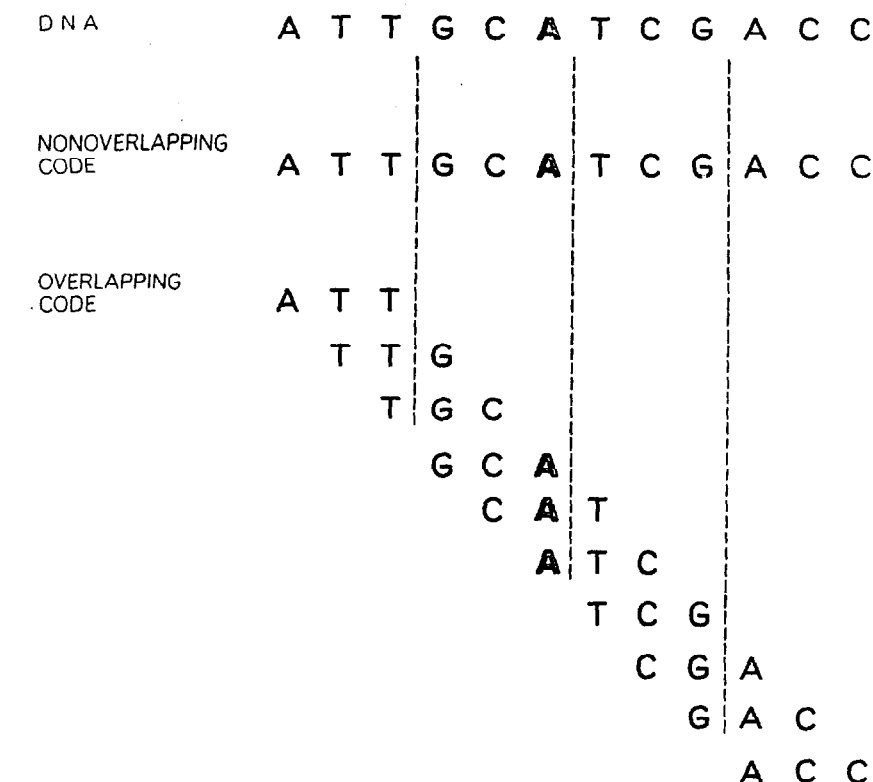
TWO CLASSES OF MUTATION result from introducing defects in the sequence of bases (A, T, G, C) that are attached to the backbone of the DNA molecule. In one class a base is simply changed from one into another, as A into G. In the second class a base is added or removed. Four bases are adenine (A), thymine (T), guanine (G) and cytosine (C).

be produced. But it now looks as if the most obvious solution is the correct one. That is, the message begins at a fixed starting point, probably one end of the gene, and is simply read three bases at a time. Notice that if the reading started at the wrong point, the message would fall into the wrong sets of three and would then be hopelessly incorrect. In fact, it is easy to see that while there is only one correct reading for a triplet code, there are two incorrect ones.

If this idea were right, it would immediately explain why the addition or the deletion of a base in most parts of the gene would make the gene completely nonfunctional, since the reading of the genetic message from that point onward would be totally wrong. Now, although our single mutations were always without function, we found that if we put certain pairs of them together, the gene would work. (In point of fact we picked up many of our functioning double mutations by starting with a nonfunctioning mutation and selecting for the rare second mutation that restored gene activity, but this does not affect our argument.) This enabled us to classify all our mutations as being either plus or minus. We found that by using the following rules we could always predict the behavior of any pair we put together in the same gene. First, if plus is combined with plus, the combination is nonfunctional. Second, if minus is combined with minus, the result is nonfunctional. Third, if plus is combined with minus, the combination is nonfunctional if the pair is too widely separated and functional if the pair is close together.

The interesting case is the last one. We could produce a gene that functioned, at least to some extent, if we combined a plus mutation with a minus mutation, provided that they were not too far apart.

To make it easier to follow, let us assume that the mutations we called plus really had an extra base at some point and that those we called minus had lost a base. (Proving this to be the case is rather difficult.) One can see that, starting from one end, the message would be read correctly until the extra base was reached; then the reading would get out of phase and the message would be wrong until the missing base was reached, after which the message would come back into phase again. Thus the genetic message would not be wrong over a long stretch but only over the short distance between the plus and the minus. By the same sort of argument one can see that for a triplet code the combination plus with plus or minus with



PROPOSED CODING SCHEMES show how the sequence of bases in DNA can be read. In a nonoverlapping code, which is favored by the author, code groups are read in simple sequence. In one type of overlapping code each base appears in three successive groups.

minus should never work [see illustration on next page].

We were fortunate to do most of our work with mutations at the left-hand end of the B gene of the rII region. It appears that the function of this part of the gene may not be too important, so that it may not matter if part of the genetic message in the region is incorrect. Even so, if the plus and minus are too far apart, the combination will not work.

Nonsense Triplets

To understand this we must go back once again to the code. There are 64 possible triplets but only 20 amino acids to be coded. Conceivably two or more triplets may stand for each amino acid. On the other hand, it is reasonable to expect that at least one or two triplets may not represent an amino acid at all but have some other meaning, such as "Begin here" or "End here." Although such hypothetical triplets may have a meaning of some sort, they have been named nonsense triplets. We surmised that sometimes the misreading produced in the region lying between a plus and a minus mutation might by chance give rise to a nonsense triplet, in which case the gene might not work.

We investigated a number of plus-with-minus combinations in which the distance between plus and minus was relatively short and found that certain combinations were indeed inactive when we might have expected them to function. Presumably an intervening nonsense triplet was to blame. We also found cases in which a plus followed by a minus worked but a minus followed by a plus did not, even though the two mutations appeared to be at the same sites, although in reverse sequence. As I have indicated, there are two wrong ways to read a message; one arises if the plus is to the left of the minus, the other if the plus is to the right of the minus. In cases where plus with minus gave rise to an active gene but minus with plus did not, even when the mutations evidently occupied the same pairs of sites, we concluded that the intervening misreading produced a nonsense triplet in one case but not in the other. In confirmation of this hypothesis we have been able to modify such nonsense triplets by mutagens that turn one base into another, and we have thereby restored the gene's activity. At the same time we have been able to locate the position of the nonsense triplet.

Recently we have undertaken one

other rather amusing experiment. If a single base were changed in the left-hand end of the B gene, we would expect the gene to remain active, both because this end of the gene seems to be unessential and because the reading of the rest of the message is not shifted. In fact, if the B gene remained active, we would have no way of knowing that a base had been changed. In a few cases, however, we have been able to destroy the activity of the B gene by a base change traceable to the left-hand end of the gene. Presumably the change creates a nonsense triplet. We reasoned that if we could shift the reading so that the message was read in different groups of three, the new reading might not yield a nonsense triplet. We therefore selected a minus and a plus that together allowed the B gene to function, and that were on each side of the presumed nonsense mutation. Sure enough, this combination of three mutants allowed the gene to function [see top illustration on page 74]. In other words, we could abolish the effect of a nonsense triplet by shifting its reading.

All this suggests that the message is read from a fixed point, probably from one end. Here the question arises of how one gene ends and another begins,

since in our picture there is nothing on the backbone of the long DNA molecule to separate them. Yet the two genes A and B are quite distinct. It is possible to measure their function separately, and Benzer has shown that no matter what mutation is put into the A gene, the B function is not affected, provided that the mutation is wholly within the A gene. In the same way changes in the B gene do not affect the function of the A gene.

The Space between the Genes

It therefore seems reasonable to imagine that there is something about the DNA between the two genes that isolates them from each other. This idea can be tested by experiments with a mutant T4 in which part of the *rII* region is deleted. The mutant, known as T4 1589, has lost a large part of the right end of the A gene and a smaller part of the left end of the B gene. Surprisingly the B gene still shows some function; in fact this is why we believe this part of the B gene is not too important.

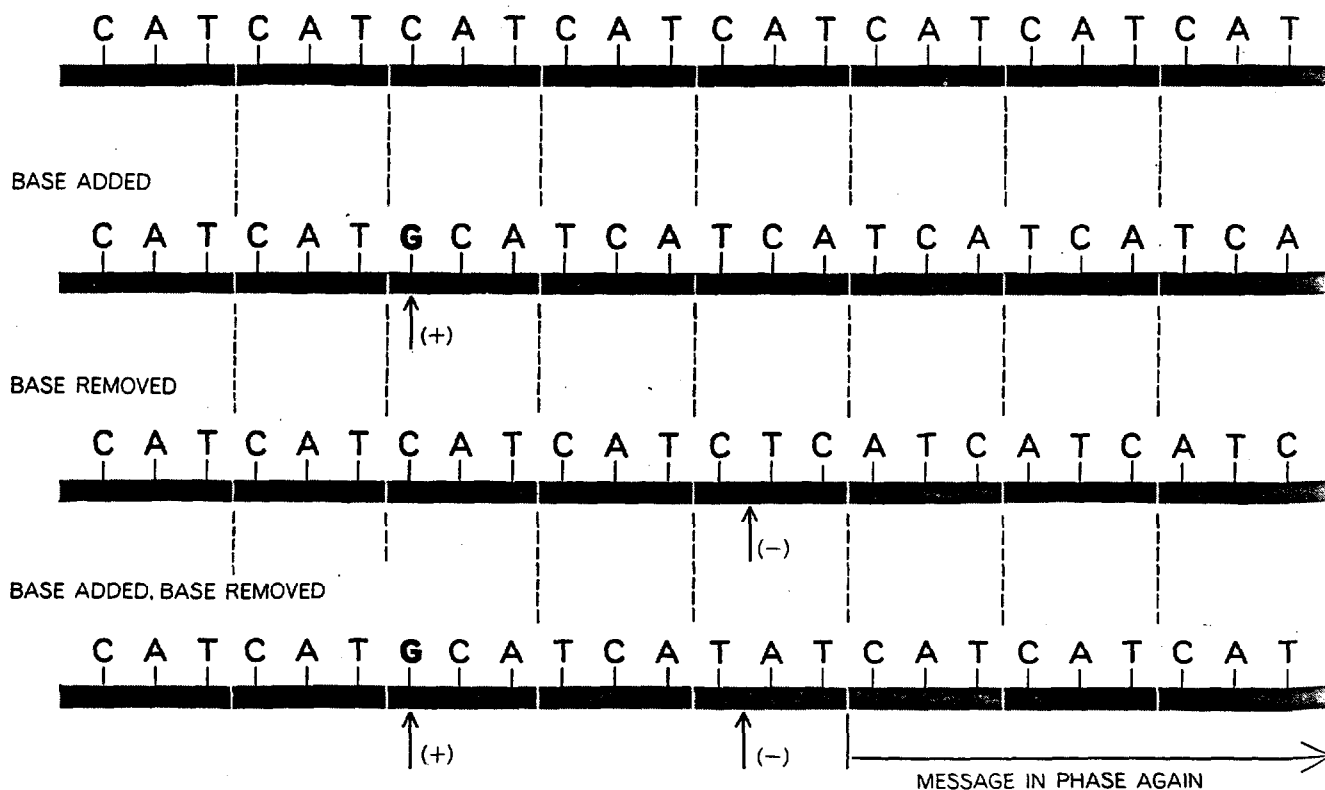
Although we describe this mutation as a deletion, since genetic mapping shows that a large piece of the genetic

information in the region is missing, it does not mean that physically there is a gap. It seems more likely that DNA is all one piece but that a stretch of it has been left out. It is only by comparing it with the complete version—the wild type—that one can see a piece of the message is missing.

We have argued that there must be a small region between the genes that separates them. Consequently one would predict that if this segment of the DNA were missing, the two genes would necessarily be joined. It turns out that it is quite easy to test this prediction, since by genetic methods one can construct double mutants. We therefore combined one of our acridine mutations, which in this case was near the beginning of the A gene, with the deletion 1589. Without the deletion present the acridine mutation had no effect on the B function, which showed that the genes were indeed separate. But when 1589 was there as well, the B function was completely destroyed [see top illustration on page 72]. When the genes were joined, a change far away in the A gene knocked out the B gene completely. This strongly suggests that the reading proceeds from one end.

We tried other mutations in the A

WILD-TYPE GENE



EFFECT OF MUTATIONS that add or remove a base is to shift the reading of the genetic message, assuming that the reading begins at the left-hand end of the gene. The hypothetical message in

the wild-type gene is CAT, CAT... Adding a base shifts the reading to TCA, TCA... Removing a base makes it ATC, ATC... Addition and removal of a base puts the message in phase again.

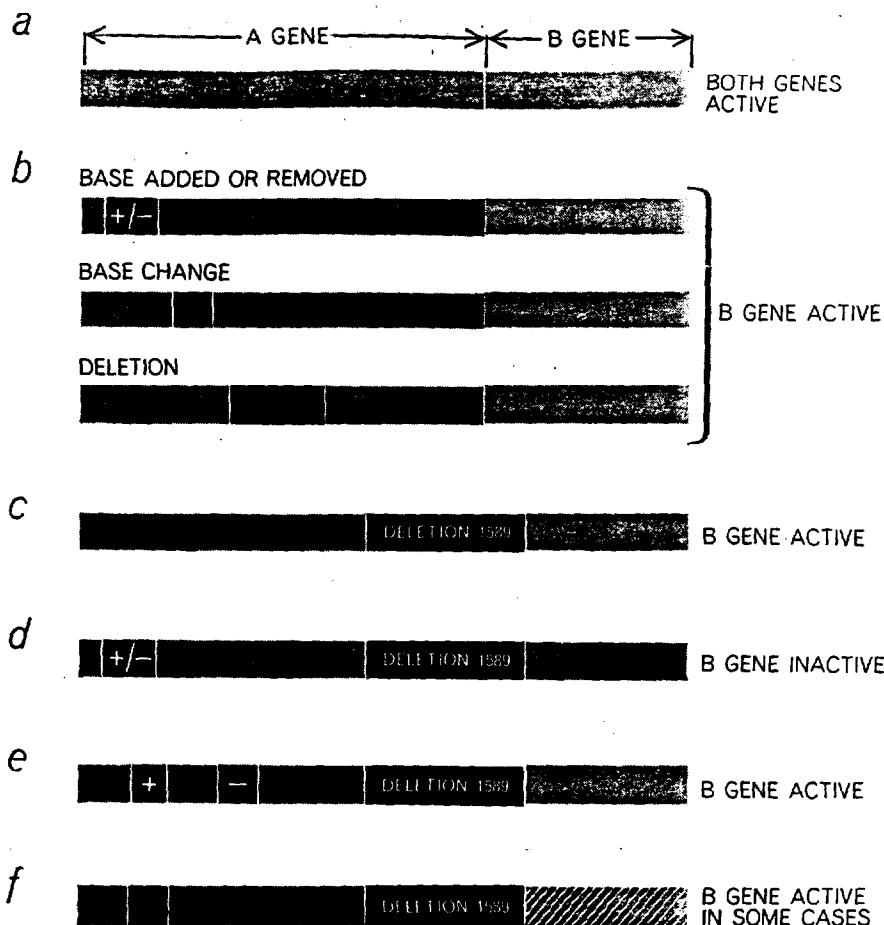
gene combined with 1589. All the acridine mutations we tried knocked out the B function, whether they were plus or minus, but a pair of them (plus with minus) still allowed the B gene to work. On the other hand, in the case of the other type of mutation (which we believe is due to the change of a base and not to one being added or subtracted) about half of the mutations allowed the B gene to work and the other half did not. We surmise that the latter are nonsense mutations, and in fact Benzer has recently been using this test as a definition of nonsense.

Of course, we do not know exactly what is happening in biochemical terms. What we suspect is that the two genes, instead of producing two separate pieces of messenger RNA, produce a single piece, and that this in turn produces a protein with a long polypeptide chain, one end of which has the amino acid sequence of part of the presumed A protein and the other end of which has most of the B protein sequence—enough to give some B function to the combined molecule although the A function has been lost. The concept is illustrated schematically at the bottom of the next page. Eventually it should be possible to check the prediction experimentally.

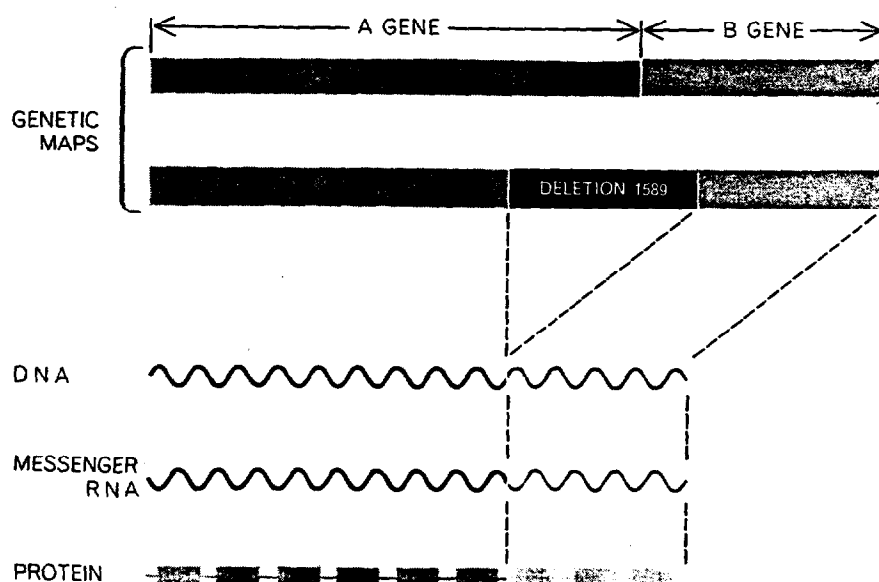
How the Message Is Read

So far all the evidence has fitted very well into the general idea that the message is read off in groups of three, starting at one end. We should have got the same results, however, if the message had been read off in groups of four, or indeed in groups of any larger size. To test this we put not just two of our acridine mutations into one gene but three of them. In particular we put in three with the same sign, such as plus with plus with plus, and we put them fairly close together. Taken either singly or in pairs, these mutations will destroy the function of the B gene. But when all three are placed in the same gene, the B function reappears. This is clearly a remarkable result: two blacks will not make a white but three will. Moreover, we have obtained the same result with several different combinations of this type and with several of the type minus with minus with minus.

The explanation, in terms of the ideas described here, is obvious. One plus will put the reading out of phase. A second plus will give the other wrong reading. But if the code is a triplet code, a third plus will bring the message back into phase again, and from then on to the end it will be read correctly. Only between



DELETION JOINING TWO GENES makes the B gene vulnerable to mutations in the A gene. The messages in two wild-type genes (a) are read independently, beginning at the left end of each gene. Regardless of the kind of mutation in A, the B gene remains active (b). The deletion known as 1589 inactivates the A gene but leaves the B gene active (c). But now alterations in the A gene will often inactivate the B gene, showing that the two genes have been joined in some way and are read as if they were a single gene (d, e, f).



PROBABLE EFFECT OF DELETION 1589 is to produce a mixed protein with little or no A-gene activity but substantial B activity. Although the conventional genetic map shows the deletion as a gap, the DNA molecule itself is presumably continuous but shortened. In virus replication the genetic message in DNA is transcribed into a molecule of ribonucleic acid, called messenger RNA. This molecule carries the message to cellular particles known as ribosomes, where protein is synthesized, following instructions coded in the DNA.

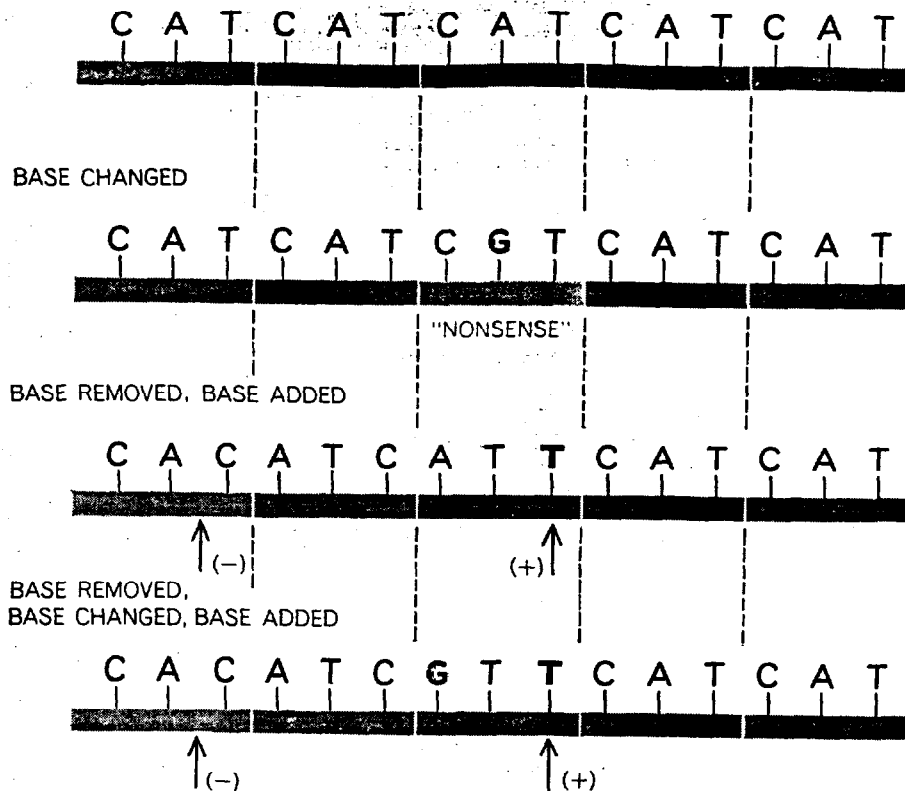
the pluses will the message be wrong [see bottom illustration on page 74].

Notice that it does not matter if plus is really one extra base and minus is one fewer; the conclusions would be the same if they were the other way around. In fact, even if some of the plus mutations were indeed a single extra base, others might be two fewer bases; in other words, a plus might really be minus minus. Similarly, some of the minus mutations might actually be plus plus. Even so they would still fit into our scheme.

Although the most likely explanation is that the message is read three bases at a time, this is not completely certain. The reading could be in multiples of three. Suppose, for example, that the message is actually read six bases at a time. In that case the only change needed in our interpretation of the facts is to assume that all our mutants have been changed by an even number of bases. We have some weak experimental evidence that this is unlikely. For instance, we can combine the mutant 1589 (which joins the genes) with medium-sized deletions in the A cistron. Now, if deletions were random in length, we should expect about a third of them to allow the B function to be expressed if the message is indeed read three bases at a time, since those deletions that had lost an exact multiple of three bases should allow the B gene to function. By the same reasoning only a sixth of them should work (when combined with 1589) if the reading proceeds six at a time. Actually we find that the B gene is active in a little more than a third. Taking all the evidence together, however, we find that although three is the most likely coding unit, we cannot completely rule out multiples of three.

There is one other general conclusion we can draw about the genetic code. If we make a rough guess as to the actual size of the B gene (by comparing it with another gene whose size is known approximately), we can estimate how many bases can lie between a plus with minus combination and still allow the B gene to function. Knowing also the frequency with which nonsense triplets are created in the misread region between the plus and minus, we can get some idea whether there are many such triplets or only a few. Our calculation suggests that nonsense triplets are not too common. It seems, in other words, that most of the 64 possible triplets, or codons, are not nonsense, and therefore they stand for amino acids. This implies that probably more than one codon can stand for one amino acid. In the jargon

WILD-TYPE GENE



NONSENSE MUTATION is one creating a code group that evidently does not represent any of the 20 amino acids found in proteins. Thus it makes the gene inactive. In this hypothetical case a nonsense triplet, CGT, results when an A in the wild-type gene is changed to G. The nonsense triplet can be eliminated if the reading is shifted to put the G in a different triplet. This is done by recombining the inactive gene with one containing a minus-with-plus combination. In spite of three mutations, the resulting gene is active.

of the trade, a code in which this is true is "degenerate."

In summary, then, we have arrived at three general conclusions about the genetic code:

1. The message is read in nonoverlapping groups from a fixed point, probably from one end. The starting point determines that the message is read correctly into groups.

2. The message is read in groups of a fixed size that is probably three, although

multiples of three are not completely ruled out.

3. There is very little nonsense in the code. Most triplets appear to allow the gene to function and therefore probably represent an amino acid. Thus in general more than one triplet will stand for each amino acid.

It is difficult to see how to get around our first conclusion, provided that the B gene really does code a polypeptide chain, as we have assumed. The second

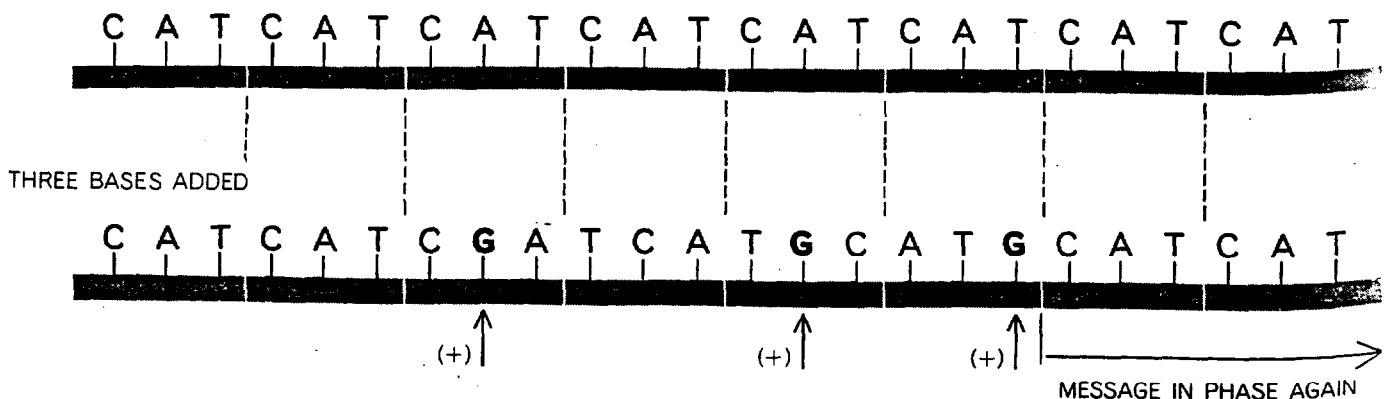
conclusion is also difficult to avoid. The third conclusion, however, is much more indirect and could be wrong.

Finally, we must ask what further evidence would really clinch the theory we have presented here. We are continuing to collect genetic data, but I doubt that this will make the story much more convincing. What we need is to obtain a protein, for example one produced by a double mutation of the form plus with minus, and then examine its amino acid sequence. According to conventional theory, because the gene is altered in only two places the amino acid sequences also should differ only in the two corresponding places. According to our theory it should be altered not only at these two places but also at all places in between. In other words, a whole string of amino acids should be changed. There is one protein, the lysozyme of the T4 phage, that is favorable for such an approach, and we hope that before long workers in the U.S. who have been studying phage lysozyme will confirm our theory in this way.

The same experiment should also be useful for checking the particular code schemes worked out by Nirenberg and Matthaei and by Ochoa and his colleagues. The phage lysozyme made by the wild-type gene should differ over only a short stretch from that made by the plus-with-minus mutant. Over this stretch the amino acid sequence of the two lysozyme variants should correspond to the same sequence of bases on the DNA but should be read in different groups of three.

If this part of the amino acid sequence of both the wild-type and the altered lysozyme could be established, one could check whether or not the codons assigned to the various amino acids did indeed predict similar sequences for that part of the DNA between the base added and the base removed.

WILD-TYPE GENE



TRIPLE MUTATION in which three bases are added fairly close together spoils the genetic message over a short stretch of the

gene but leaves the rest of the message unaffected. The same result can be achieved by the deletion of three neighboring bases.